Chapter 30
Risk-Based Privacy-Aware Information Disclosure

Alessandro Armando
Fondazione Bruno Kessler, Italy & University of Genova, Italy

Michele Bezzi
SAP Labs – Sophia-Antipolis, France

Nadia Metoui
Fondazione Bruno Kessler, Italy & University of Trento, Italy

Antonino Sabetta
SAP Labs – Sophia-Antipolis, France

ABSTRACT

Risk-aware access control systems grant or deny access to resources based on the notion of risk. It has many advantages compared to classical approaches, allowing for more flexibility, and ultimately supporting for a better exploitation of data. The authors propose and demonstrate a risk-aware access control framework for information disclosure, which supports run-time risk assessment. In their framework access-control decisions are based on the disclosure-risk associated with a data access request and, differently from existing models, adaptive anonymization operations are used as risk-mitigation method. The inclusion of on-the-fly anonymization allows for extending access to data, still preserving privacy below the maximum tolerable risk. Risk thresholds can be adapted to the trustworthiness of the requester role, so a single access control framework can support multiple data access use cases, ranging from sharing data among a restricted (highly trusted) group to public release (low trust value). The authors have developed a prototype implementation of their framework and have assessed it by running a number of queries against the Adult Data Set from the UCI Machine Learning Repository, a publicly available dataset that is widely used by the research community. The experimental results are encouraging and confirm the feasibility of the proposed approach.

DOI: 10.4018/978-1-5225-7113-1.ch030

INTRODUCTION

The increase in the amount of data generated by today's digital society is astonishing. According to IDC estimate (IDC, 2014), the global volume of digital data will double every two years, reaching 44 trillion gigabytes by 2020. Potentially organizations are now in the position to fully exploit this huge amount of diverse datasets to create new data-based businesses as well as optimizing existing process (e.g., real-time customization). On the other hand, often, organizations are not fully leveraging this potential due to the lack of appropriate dissemination and control mechanisms, which supports risk-based decision making, balancing the advantages of information access with the security. Personal information is particularly critical, since they are subject to strict regulations, and enterprises will have to comply with it to avoid significant fines and impact on reputation. As a result, most organizations strongly limit (even internally) the sharing and dissemination of data making most of the information unavailable to decision-makers and therefore do not exploit the power of existing data sources.

Already a few years ago, the JASON report (JASON, 2004) indicated that the inflexibility of existing access control mechanisms is a major obstacle with dealing with diverse data sources in dynamic environments. To address this issue, access control mechanisms based on risk estimation (i.e., risk-aware access control) have been put forward (Cheng, 2007). In a nutshell, in risk-aware access control access control decisions are based on an estimation of expected cost and benefits and only not (as in traditional access control systems) on a policy statically defining stating which requests should be allowed and which should be denied. In a risk-aware access control system, for each access request, the corresponding risk is estimated and compared with a risk-threshold. If the risk is less than a given risk threshold, then access is granted, otherwise it is denied. This allows for a better exploitation of the data than in traditional access control system while controlling risk. Although existing risk-aware access control models enjoy many advantages and allow for a better management and exploitation of the data, they suffer from a number of drawbacks that limit its effectiveness. For instance, most existing risk-based access control models only support binary access decision (i.e., the outcome is either allowed or denied), whereas in real-life we often have exceptions based on additional conditions (e.g., I cannot disclose these data, because they contain personal identifiable information, but I can disclose an anonymized version of the data). In other words, the system should enforce appropriate risk mitigation measures, and relevant part of additional information could be shared. From a risk perspective, such mitigation measures have the effect of decreasing the risk associated with the release of the data.

Anonymization is a commonly used practice to reduce privacy risk, obfuscating, in part or completely, the personal identifiable information in a dataset. Anonymization methods include (Ciriani, 2009): suppressing part of or entire records; generalizing the data, i.e., recoding variables into broader classes (e.g., releasing only the first two digits of the zip code) or rounding/clustering numerical data; replacing identifiers with random values (e.g., replacing a real name with a randomly chosen one). To quantify the level of anonymity, several metrics have been proposed in the literature (see (Bezzi, 2010; Clifton, 2013) for a review). These metrics differ in a number of ways, but they all express the risk of disclosing personal-identifiable information when releasing a given dataset. Anonymization increases protection, by lowering the privacy risk, and enables a wider exploitation of the data, but it clearly impacts the utility of the data. Accordingly, different level of anonymization should be considered depending on a number of factors, often known at run-time only, such as the trustworthiness of the requester or security context of the query.